New Apex in Proteome Analysis

Ly, Tony; Lamond, Angus I.

Published in:
Cell Systems

DOI:
10.1016/j.cels.2017.06.009

Publication date:
2017

~~Seeing the big picture: fFaster and Ddeeper Pproteome Aanalysis~~

New Apex in Proteome Analysis

~~[AU: Original title too long. I have suggested two alternatives, with a slight preference for the second one because it is "snappier" , conveys that this work represents a new high-water mark for the field, and is less redundant with the Summary.]~~

Tony Ly and Angus I. Lamond

[1]Centre for Gene Regulation and Expression, University of Dundee, Dundee, DD1 5EH, United Kingdom

[*]Correspondence: ailamond@dundee.ac.uk

SUMMARY

~~Olsen and colleagues show i~~Improved sample processing workflows enable~~for~~ rapid~~ly obtaining comprehensive~~, deep analysis of human cellular and tissue proteomes. ~~[AU: OK?]~~

~~OPENING PARAGRAPH~~

Technological advances in chromatography, mass spectrometry (MS), and data analysis have heralded a new era of proteomics that is characterised by increasingly detailed analyses of cellular proteomes from human cells and model organisms. I~~n an article i~~n this issue ~~of Cell Systems~~, Olsen and colleagues describe how they have improved significantly upon "~~"~~classic"~~"~~ workflows to allow rapid and impressively deep proteome coverage (~584,000 unique peptide sequences corresponding to 14,200 proteins identified) in a single human cancer cell line (HeLa) (Bekker-Jensen et al. 2017). The depth achieved in this study is indeed comparable with~~ ~~or even better~~ ~~than~~ the two recently reported "~~"~~drafts"~~"~~ of the entire human proteome (Ahmad et al. 2014; Mollenhauer et al. 2014), which are based on data spanning diverse cell and tissue types. The deep coverage reported by Olsen and colleagues was enabled by the ~~using~~ use of multiple digestion enzymes and extensive peptide pre-fractionation, prior to LC-MS/MS.

A fast, high-throughput, and quantitative assay measuring the proteome of a single population of human cells in depth, either in tissue culture~~,~~ or in tissues, would greatly accelerate biomedical research. For example, this could~~an~~ provide an objective and rigorous molecular definition of cell type identity (Hukelmann et al. 2016; Geiger et al. 2012; Ly et al. 2014). Importantly, it could~~an~~ ~~potentially~~ also enable comprehensive

“~~“~~snapshots”~~”~~ of proteome remodelling during diverse cellular processes, such as cell division ~~(Ly et al. 2014)~~ and differentiation ~~(Van Hoof et al. n.d.)~~.

Making proteome analyses “~~“~~comprehensive”~~”~~ poses major analytical challenges. One problem, for example, is that protein expression levels in human cells typically span a dynamic range of 7~~–~~–8 orders of magnitude. Unlike the situation for nucleic acid analysis, where the detection of low abundance species is facilitated by PCR amplification, there is no comparable amplification process available to aid detection of low abundance peptides. Another challenge is that searching reference databases, as required for peptide identification in most MS-based proteomics workflows, can be overwhelmed if the full diversity of potential peptide analyte species is included. For example, the human genome encodes at least 42,210 protein isoforms (SwissProt reviewed UniProt reference proteome), potentially giving rise to millions of distinct peptides upon protease digestion. The diversity of peptide analytes is further increased by both genetic variation and via post-translational modifications (PTMs).~~,~~ T~~t~~he latter ~~hugely~~ massively increas~~es~~~~ing~~ the size of the database ~~size~~ needed to describe the full set of possible modified peptides. Furthermore, some PTMs are technically difficult to detect, either because the modifications are labile and easily lost during sample handling and mass spectrometry~~& MS~~, or because they are present in low abundance and/or low stoichiometry.

Given these challenges, the depth of proteome coverage achieved by Olsen and colleagues, including detection of ~7,000 acetylation sites and ~10,000 phosphorylation sites, is a remarkable analytical achievement. To a large extent, this technical advance represents the effective combination of several methodological improvements previously

described individually in the literature, including~~: i.e.,~~ extensive peptide pre-fractionation (46 fractions by high pH reverse phase)(Wang et al. 2011), use of short LC-gradients (30 minutes) (Kelstrup et al. 2012~~; Hsieh et al. 2012~~), and protein digestion using multiple proteases (trypsin, LysC, chymotrypsin and GluC)(Swaney et al. 2010). In this study, ~34~~–~~35 hr of MS instrument time was sufficient to analyse by LC-MS/MS 46 fractions of fractionated tryptic peptides from a digest of total HeLa cell extract, resulting in identification of 0.17 million unique peptide sequences. Including the data from the three other protease digests analysed resulted in identification of another ~0.25 million peptides, resulting in a combined sum of 0.42 million unique peptide sequences. By several measures, including near-complete coverage of a "~~"~~core"~~"~~ set of human protein complexes (CORUM), the authors argue that their analysis provides the~~,~~ "essentially complete HeLa proteome." ~~.~~

So, taking this claim at face value, how close are we to having a "~~"~~*complete proteome*"~~"~~ of a human cell line?

Answering this question involves agreeing first of all ~~agreeing~~ on a definition of what "~~"~~complete"~~"~~ actually means (Mann et al. 2013). While a detailed discussion of such a definition is beyond the scope of this Preview, the study by Olsen and colleagues, along with other recent studies with high proteome depth, raise interesting points relevant to refining what is understood by "~~"~~complete~~?~~." For example, the combined number of peptides identified here, with six~~6~~ technical replicates using trypsin digestion (0.36 million peptides), represents ~19% of the theoretical estimate of ~1.9 million peptides, i.e. the *in silico* tryptic digestion of the protein sequences encoded by the estimated >12,000 open reading frames annotated in the reference human genome. Furthermore, the

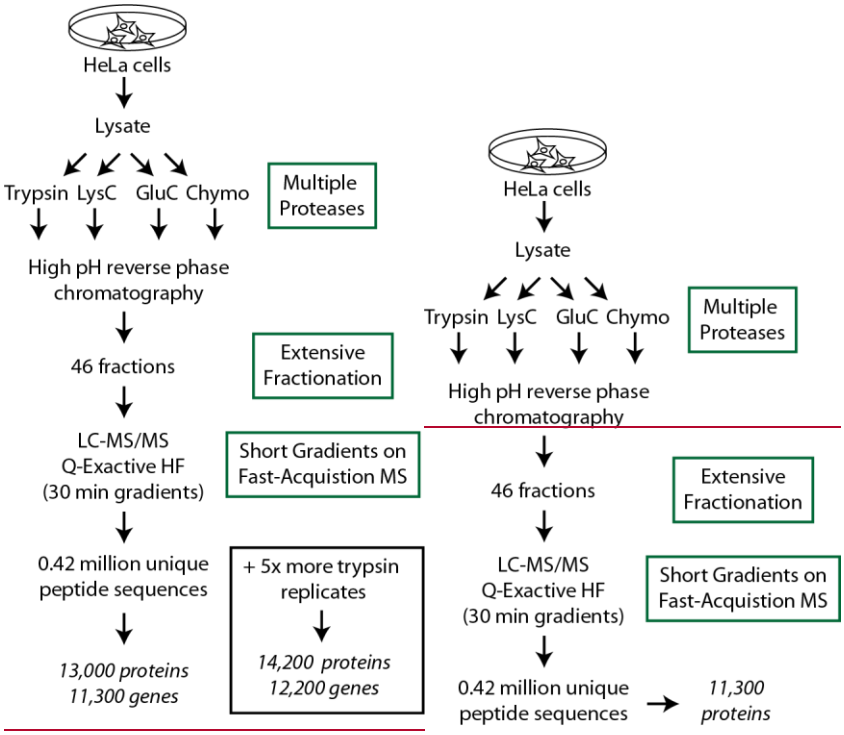
average protein sequence coverage reported here (~52%), while potentially the highest achieved thus far for a single cell line, nonetheless means, by definition, that ~48% of protein sequences in this cell line were not detected. The authors also point out that their phosphorylation dataset, covering ~10,000 phosphorylation sites, while remarkably detailed considering that they did not use phosphopeptide enrichment, does not include a wide range of phosphopeptides previously identified in HeLa and other similar human cancer cell lines. Many other biologically important PTMs were not analysed in this study. In practice, therefore, the missing information places limitations on the resolution of the proteome with respect to isoform expression.

This question of "~~"~~proteome completeness"~~"~~, is reminiscent of a phenomenon in mathematics called the "~~"~~*coastline paradox*,"~~"~~; i.e. in which~~,~~ the measured value for the length of a convoluted coastline continually increases as the resolution of the measurement increases ~~(Mandelbrot 1967)~~. Extending the metaphor, just as the infinitely detailed and changing nature of a coastline's contour necessitates drawing a practical limit on what is geographically meaningful, it will be important in future to evaluate, *with respect to biological function*, what depth of protein and PTM coverage will not only be practical to measure, but also required to mechanistically describe the phenotype of the cell or ~~/~~tissue under study. Therefore, for the time being, we would respectfully suggest avoiding the potentially contentious term "~~"~~complete"~~"~~ and focus instead on "~~"~~comprehensive"~~"~~ proteome measurements.

Semantics aside, Olsen and colleagues clearly present here a benchmark study that showcases the hugely impressive depth of coverage that can now be achieved by MS-based proteomics. Using a combination of extensive peptide pre-fractionation, short

LC-gradients and protein digestion with multiple proteases, the authors obtained a comprehensive dataset profiling the HeLa cell proteome in a relatively short timeframe. The same workflow was also shown to be useful for obtaining comprehensive proteomic datasets on other cell lines and patient tissue samples. We look forward ~~now~~ to seeing this workflow applied in future studies on other cell types and used to take comprehensive ~~"~~"~~snapshots~~"~~" of cellular and tissue proteomes in flux.

Figure 1. A workflow for comprehensive proteome analysis using multiple proteases, extensive peptide fractionation, and short LC-MS/MS gradients. The cumulative depth obtained from HeLa is 14.2k proteins (12.2k protein-coding genes) using the above workflow and five additional technical repeats of trypsin. The impressive analytical depth achieved by Olsen and colleagues (Bekker-Jensen et al. 2017) represents the effective combination of several methodological improvements previously described individually in the literature, but which, to our knowledge, have not been used together before. It should be possible for other groups to readily adopt the workflows. ~~**[AU: OK? Please ensure that all of the technical details in the figure are consistent with the text. For example, the figure states 11,300 proteins whereas the first paragraph states 14,200 proteins identified. Also, I suggest putting the number of proteins at the bottom of the workflow, instead of off to the side, because it is currently not clear why the number of proteins is vertically aligned with the methodological improvements]**~~



REFERENCES

~~[AU: Max of 10 references allowed, including the paper you are Previewing, and unfortunately we must be quite strict about the limit. Could I ask you to please trim three references?]~~

Ahmad, S. et al., 2014. A draft map of the human proteome. *Nature*, 509(7502), pp.575–581.

Bekker-Jensen, D.B., et al., 2017. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. Cell Systems, XX, XX–XX.

Geiger, T. et al., 2012. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Molecular & Cellular Proteomics*, 11(3).

~~Hsieh, E.J. et al., 2012. Effects of Column and Gradient Lengths on Peak Capacity and~~

~~Peptide Identification in Nanoflow LC-MS/MS of Complex Proteomic Samples. *Journal of the American Society for Mass Spectrometry*, 24(1), pp.148–153.~~

Hukelmann, J.L. et al., 2016. The cytotoxic T cell proteome and its shaping by the kinase mTOR., 17(1), pp.104–112.

Kelstrup, C.D. et al., 2012. Optimized Fast and Sensitive Acquisition Methods for Shotgun Proteomics on a Quadrupole Orbitrap Mass Spectrometer. *Journal of Proteome Research*, 11(6), pp.3487–3497.

Ly, T. et al., 2014. A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *eLife*, 3, p.e01630.

~~Mandelbrot, B., 1967. How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. *Science*, 156(3775), pp.636–638.~~

Mann, M. et al., 2013. The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Molecular Cell*, 49(4), pp.583–590.

Mollenhauer, M. et al., 2014. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), pp.582–587.

Swaney, D.L., Wenger, C.D. & Coon, J.J., 2010. Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. *Journal of Proteome Research*, 9(3), pp.1323–1329.

~~Van Hoof, D. et al., Phosphorylation Dynamics during Early Differentiation of Human Embryonic Stem Cells. *Cell Stem Cell*, 5(2), pp.214–226.~~

Wang, Y. et al., 2011. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *PROTEOMICS*, 11(10), pp.2019–2026.